The Effect of Per-Student Expenditure on Percentage of Students Satisfactory in Mathematics

Abhinav Tirath

March 22, 2021

ECON 3161: Dr. Shatakshee Dhongde

Abstract

This study investigates whether per-student expenditure is correlated with the quality of education, using the percentage of students satisfactory in mathematics as an analog. Although the simple linear regression model did not find a significant relationship, all multiple regression models indicated that the correlation was significant and positive. In addition, the two variables unique to this study, *read4* and *lunch*, were the most important secondary independent variables. However, more practical conclusions can be made when models do not include *read4*. A model with only per-student expenditures and *lunch* was found to explain a large amount of variation in our dependent variable. Overall, we conclude that increasing expenditures per student can increase the percentage of students satisfactory in 4th grade math and the overall quality of education at a school.

I. Introduction

It is undeniable that the quality of education is extremely important for a variety of reasons. On an individual basis, a high-quality education fosters stability and financial security. It also encourages upward mobility and self-dependency. As a whole, education helps prepare children for life. As a result, a higher quality education will help students have a better life.

On a national level, the quality of education is also extremely important; it helps determine the levels of economic growth and poverty. Hanushek, Ruhose, and Woessmann (2015) found that "differences in the quality account for 20-30 percent of the state variation in per-capita GDP." An article by the UN revealed that "one extra year of education is associated with a reduction of the Gini coefficient by 1.4 percentage points" (Quality Education: Why It Matters). It is clear that governments should be striving to increase education quality, as it has a profound effect on both the economy and poverty rates.

For these reasons and more, the UN has made inclusive and equitable quality education one of its seventeen sustainable development goals. The UN has found, however, that the "inequalities in education are exacerbated due to Covid-19" (Goal 4 | Education). It is our responsibility as a society now more than ever to identify how to close these gaps. Before we are able to take action to increase equality, we must first determine what will lead to the highest quality of education for students.

We believe that expenditures per student (EPS) is a key factor in determining quality of education, and this paper aims to investigate this relationship. Even in America, large funding inequities exist between public schools. Mcfarland et al. (2019) found that the richest quarter of schools paid $473 more per student than the poorest quarter. Overall, a disparity in school funding may increase inequality and lead to a lower quality of education for certain students.

But quality of education is difficult to quantify precisely. In this paper, we use the percentage of students satisfactory in fourth grade mathematics (PSM) to estimate the quality of education.

We hypothesize that there will be a positive correlation between EPS, our primary independent variable, and PSM, our dependent variable. The rationale is that the additional funding will be used to improve the quality of each student's understanding of the material. For example, the funds could be used to hire more teachers to decrease the teacher-student ratio, allowing for more personalized learning and one-on-one time. Another example is that the school will be able to purchase new or higher quality resources, such as textbooks or interactive software, which may allow students to grasp content more easily. Overall, we believe an increase in EPS will result in an increase in PSM.

This paper will quantify and estimate the key relationship through the lens of single and multiple regression analysis.

II. Literature Review

We reviewed several economic journal papers before beginning our analysis on this topic. The three we discuss below are the most relevant.

Mutuku and Korir (2019) explore a variety of variables related to government expenditure that affect student performance in Kenyan schools. The Kenyan government implemented a program to increase government funding of primary schools in 2003. The authors wanted to examine the real effects that this funding has had. Overall, they found several key factors related to government expenditure that were positively correlated and statistically significant with respect to a student's performance on exams. These include the number of classes, teachers, books, and availability of toilets. One additional note they included was that school distance from amenities was negatively correlated with student performance. Importantly, they confirmed teacher characteristics and experience were a major factor in determining the quality of the school and the student test scores. Finally, they included some policy recommendations. They suggested increasing government expenditure on schools in general. They also recommended that these additional funds be put towards improving the quality of classes and teachers.

Carpenter (2006) conducted a study analyzing South Carolina elementary schools with three primary objectives. First, they wanted to determine the relationship between school size and academic achievement on statewide exams. After analysis, however, no relationship was found between the two factors when controlling for socioeconomic student status. Second, the study investigated how related school size was to the operational cost per pupil, which is the same as EPS. This relationship was found to significant and negative, implying that there is a much lower EPS for larger schools. This result is expected, as there is a large fixed cost for each public school. Third and most importantly, the study used multiple regression analysis to study whether academic achievement could be predicted by school size, EPS, and the socioeconomic status of the student body. It found that test scores were not predicted by school size or operating cost per pupil. Rather, it concluded that only socioeconomic status was the only predictor. The paper recommends that the factors that help children achieve academic success must be explored further.

Briggs (2013) conducted a study on the effect of school size and per-pupil expenditure on a standardized test called the Terra Nova Assessment in Catholic elementary schools in America. More specifically, the study aimed to determine how the variance in the assessment scores could be accounted for by the variance in these factors. The study employed correlation and regression analysis to examine relationships. The study determined that the relationship was not substantive enough to be used to make practical decisions about the future of Catholic education. As a result, the author concluded that decisions the pertain to EPS and school size can be made without major concern about the effects on student achievement.

Our paper contributes in a variety of ways. First, it can serve to substantiate or refute the assertion that EPS is not a reliable predictor of academic achievement and quality of education in the United States. Of the studies above, the only one that determined that increased expenditure led to higher test scores was the Kenyan study. On the other hand, both American studies concluded that EPS does not affect test scores significantly. Still, this conclusion is difficult to grapple to completely accept. It seems obvious that additional resources should lead to better educational opportunities. Since our dataset is a very different subset than the previous two studies, our findings can help determine whether the conclusions of the previous American studies apply to schools across America. Second, our paper introduces two new variables as potential predictors, the percentage of students eligible for free or reduced lunch and the percentage of students who are satisfactory in reading. Regarding the former, neither study listed above has used this factor in multiple linear regression. Controlling for this factor may help us determine that EPS does have a measurable and statistically significant impact on test scores. Regarding the latter, no previous study has used one test score to help predict another. It will be interesting to see what kind of impact controlling for this variable will have on the coefficient for ESP in our regression analysis.

III. Data

We obtained our dataset from the meap01.dta dataset accompanying Introductory Econometrics: A Modern Approach, Jeffrey M. Wooldridge. The dataset was compiled by the Michigan Department of Education. We have 1,823 observations in the dataset, and no variables are missing from any observation. The 2001 dataset measures the score for students from 1,823 schools on the Michigan Educational Assessment Program (MEAP). There are several variables in the dataset: number of students enrolled, percentage of students eligible for free or reduced lunch, percentage of students satisfactory in reading, percentage of students satisfactory in mathematics, total spending, and EPS. In addition, the logarithmic form of three variables are included: the number of students enrolled, total spending, and EPS. Finally, the district and building code are included. They are used to identify schools, but we will eliminate them from the dataset because they have no bearing on our key relationship. No other adjustments have been made to the data. Descriptions for each variable can be found in Table 1.

Table 1 – Variable Descriptions

Name	*Description*
math4 (dependent variable)	percentage of students satisfactory in 4th grade math
read4	percentage of students satisfactory in 4th grade reading
lunch	percentage of students eligible for free or reduced lunch
enroll	number of students enrolled at the school
expend	total school expenditure in $
exppp	expenditure per pupil in $
lexpend	natural logarithmic form of expend
lexppp (primary independent variable)	natural logarithmic form of exppp

As discussed above, we want to measure the effect of the EPS on the percentage of students satisfactory in mathematics. Clearly, our dependent variable is math4. For our primary independent variable, we have two options: exppp and lexppp. We chose to go forward with lexppp. This will more easily allow us to extrapolate meaningful conclusions because we will be able to determine how a percentage change in EPS affects PSM. As a result, all variables other than exppp, lexppp, and math4 will be used as control variables. The scatterplot of our dependent variable and independent variable in Figure 1 reveals that the two variables appear to be slightly positively correlated at a first glance. However, the plot has a lot of clustering, and the relation may not be linear. In addition, we can determine from the outliers that the EPS is not perfectly correlated with PSM. The samples with the lowest values of lexppp do not have the lowest scores, and the highest values of lexppp do not have the highest scores.

Figure 1 – Scatterplot of the Natural Logarithm of EPS on PSM

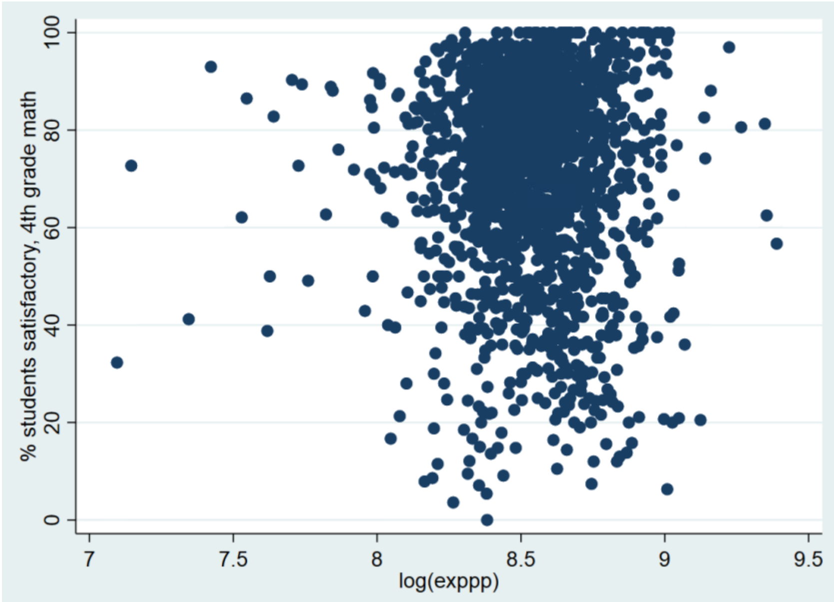

Table 2 provides summary statistics for each relevant variable.

Table 2 – Variable Descriptive Statistics

Variable	*Mean*	*Std. Dev.*	*Min*	*Max*
math4	71.91	19.95	0	100
read4	60.06	19.15	0	100
lunch	39.25	26.42	0	100
enroll	401.93	169.83	62	1496
expend	2036984	864936.1	275985	7665998
lexpend	14.44	0.412	12.53	15.85
lexppp	8.53	0.21	7.10	9.39

We must evaluate whether our regression models meet the Gauss Markov Assumptions. Without these assumptions, we are not able to confirm the unbiasedness of our ordinary least squares (OLS) model or homoskedasticity. The five assumptions are explained below:

1. Linear in Parameters: When constructing our regression models for Section IV, we ensured that all of our parameters will be linear. As a result, this assumption is validated.

2. Random Sampling: The data was sourced from the Michigan Department of Education, but the organization does not clarify which schools provided data. It is unclear whether the dataset was obtained through random sampling. We must assume that this is true to determine that our regression is unbiased.
3. No Perfect Collinearity: There are two components to perfect collinearity. First, we must ensure that no independent variables are constant. From Table 2, no standard deviations are 0. Therefore, our first part is true. Second, we must check whether there is an exact linear relationship among any of our control variables. As we can see in Table 3, no correlation coefficient is 1, so both parts of our assumption of no perfect collinearity have held. We do note, however, that expend and enroll have high coefficients, so this is a possible case for multicollinearity.

Table 3 – Correlation Coefficients Between Independent Variables

	read4	lunch	enroll	expend	lexpend	lexppp
read4	1					
lunch	-0.61	1				
enroll	-0.11	-0.01	1			
expend	-0.14	0.13	0.82	1		
lexpend	-0.10	0.05	0.81	0.94	1	
lexppp	-0.02	0.22	-0.31	0.24	0.22	1

4. Zero Conditional Mean: This means that the expected value of the error term, which is denoted by u in this paper, is 0 given any values of the independent variables. There are likely other important factors not included in the dataset that correlate with the independent variables and affect PSM. One example is the quality of the teachers. As a result, although we will assume that this is true, it is a difficult assumption to make.
5. Homoskedasticity: We must assume that the variance of error term is the same given any values of our explanatory variables. The values of our independent variables most likely contain some information about the variability of unobserved factors. As a result, this assumption is not able to be confirmed.

Since it is not clear that all of our assumptions can be made clearly, we will interpret all results carefully.

IV. Results

Model 1:

Model 1 is a simple linear regression between our primary independent variable, the natural logarithm of EPS, and our dependent variable, PSM. Our model equation is:

$$math4 = B_0 + B_1 * log(exppp) + u$$

We used STATA to calculate the values of B_0 and B_1, and this can be seen in Figure 2 below.

Figure 2 – Stata Output for Model 1

. regress math4 lexppp

Source	SS	df	MS
Model	193.269337	1	193.269337
Residual	725264.781	1,821	398.278298
Total	725458.05	1,822	398.16578

Number of obs	=	1,823
F(1, 1821)	=	0.49
Prob > F	=	0.4861
R-squared	=	0.0003
Adj R-squared	=	-0.0003
Root MSE	=	19.957

math4	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lexppp	-1.515052	2.1749	-0.70	0.486	-5.780612	2.750509
_cons	84.83715	18.56462	4.57	0.000	48.42697	121.2473

This estimated equation is as follows:

$$math4 = 84.837 - 1.515 * log(exppp)$$

The results of this model are highly surprising, and it does not conform to our initial hypothesis. The results indicate that a 1% increase in expenditures per student results in a 1.515% decrease in the percentage of students satisfactory in 4^{th} grade mathematics. We expected the coefficient for $log(exppp)$ to be positive.

However, almost none of the variance of math4 is actually explained by $log(exppp)$, so this equation does not provide much information regarding the relationship. This is true because the R-squared value is 0.0003 and the adjusted R-squared value is -0.003. These values are trivial. This does not mean, however, that our primary independent variable is insignificant. It means that we are not controlling for the unobserved values sufficiently. In our next models, we will use multiple linear regression so that we can control for many more variables. This will likely give us more insight into the relationship between our primary independent and dependent variables.

Model 2:

Model 2 uses our primary independent variable and all level independent variables. Model 2 is represented as:

$$math4 = B_0 + B_1 * log(exppp) + B_2 read4 + B_3 lunch + B_4 enroll + B_5 expend + u$$

Our STATA output is listed in Figure 3:

Figure 3 – Stata Output for Model 2

```
. regress math4 lexppp read4 lunch enroll expend

      Source |       SS           df       MS      Number of obs   =     1,823
-------------+----------------------------------   F(5, 1817)      =    948.84
       Model |  524555.886         5  104911.177   Prob > F        =    0.0000
    Residual |  200902.164     1,817   110.56806   R-squared       =    0.7231
-------------+----------------------------------   Adj R-squared   =    0.7223
       Total |   725458.05     1,822   398.16578   Root MSE        =    10.515

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lexppp |   6.806119    2.91684     2.33   0.020     1.085406    12.52683
       read4 |   .7828376   .0165877    47.19   0.000     .7503047    .8153705
       lunch |    -.10646   .0122221    -8.71   0.000     -.130431   -.0824891
      enroll |   .0076482   .0062959     1.21   0.225    -.0046997    .0199961
      expend |  -2.25e-06   1.21e-06    -1.86   0.063    -4.63e-06    1.25e-07
       _cons |  -27.49358   24.84519    -1.11   0.269    -76.22171    21.23455
------------------------------------------------------------------------------
```

The estimated equation for Model 2 is as follows:

$$math4 = -27.494 + 6.806 * log(exppp) + .783 read4 - .107 lunch + .008 enroll$$

$$-(2.25 * 10^{-6}) expend$$

The results of this model are much more in line with what we had expected. The coefficient for log(exppp) is 6.806. This indicates that a 1% increase in expenditures per pupil increases the percentage of children satisfactory in mathematics by 6.806%. As our hypothesis predicted, the sign is of the coefficient is positive. In addition, this model is able to explain the much of the variation in math4. We know this to be true because our R-squared value is 0.723. The adjusted R-squared value is 0.722. Even when we account for the number of predictors in our model, the model is able to explain the variation in math4 effectively.

Comparing this to our simple linear regression model, there are two primary differences. First, it is clear that coefficient for $log(exppp)$ in the two models are very different. Whereas Model 1 predicted a decrease of PSM with an increase in EPS, Model 2 predicted the opposite. The magnitude of the effect on math4 with a 1% change in EPS is also much different; Model 2 predicted the effect to be about 4 times greater. Second, the magnitude of the R-squared and adjusted R-squared values are much greater in Model 2 than in Model 1. As a result, we determine that the variation in math4 is much better predicted by Model 2 than by Model 1. Overall, the results of Model 2 match our hypothesis more closely, and the R-squared value informs us that it is a better predictor of math4.

Model 3:

Instead of using level variables from Model 2, Model 3 uses the log form for total expenditure and does not use enroll. Resultingly, Model 3 uses our primary independent variable, lexpend, read4, and lunch as explanatory variables. Model 3 is represented as:

$$math4 = B_0 + B_1 * log(exppp) + B_2 read4 \; + B_3 lunch \; + B_4 lexpend + u$$

Our STATA output is listed in Figure 4:

Figure 4 – Stata Output for Model 3

```
. regress math4 lexppp read4 lunch lexpend

      Source |       SS           df       MS      Number of obs   =     1,823
-------------+----------------------------------   F(4, 1818)      =   1182.22
       Model |  524005.833         4  131001.458   Prob > F        =    0.0000
    Residual |  201452.217     1,818    110.8098   R-squared       =    0.7223
-------------+----------------------------------   Adj R-squared   =    0.7217
       Total |   725458.05     1,822   398.16578   Root MSE        =    10.527

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lexppp |   3.383528   1.222901     2.77   0.006     .9850887    5.781967
       read4 |   .7853037   .0165835    47.35   0.000      .752779    .8178284
       lunch |  -.1079759   .0122703    -8.80   0.000    -.1320412   -.0839106
     lexpend |  -1.273562   .6187864    -2.06   0.040    -2.487169   -.0599549
       _cons |   18.50375   12.00487     1.54   0.123     -5.04104    42.04853
------------------------------------------------------------------------------
```

The estimated equation for Model 3 is as follows:

$$math4 = 18.504 + 3.384 * log(exppp) + .785 read4 - .108 lunch - 1.274 expend$$

In Model 3, the coefficient for $log(exppp)$ is positive. The model indicates that a 1% increase in expenditures per pupil increases the percentage of children satisfactory in mathematics by 3.384%. Moreover, the R-squared value and adjusted R-square value are both 0.72.

The coefficient of for $log(exppp)$ and R-squared values between Model 2 and Model 3 are similar. Both models suggest that the effect of a 1% increase in EPS will result in a positive change in

PSM. However, the coefficient for $log(exppp)$ in Model 3 is less than half than that of Model 2. Model 3 predicts that the effect of a 1% change in EPS is significantly smaller than does Model 2. The adjusted R-squared and R-squared values are almost identical in both equations. It is difficult to determine whether the predictions in Model 2 or Model 3 are more reliable.

Model 4:

Going forward, we will eliminate the *read4* variables from our model. Although it is interesting to evaluate models with the variable, we will be able to make more practical implications when we are not using one metric of academic achievement to predict another. Model 4 is exactly like Model 2, except read4 is excluded. Model 4 is represented as:

$$math4 = B_0 + B_1 * log(exppp) + B_2 lunch\ + B_3 enroll\ + B_4 expend + u$$

Our STATA output is listed in Figure 5:

Figure 5 – Stata Output for Model 4

```
. regress math4 lexppp lunch enroll expend
```

Source	SS	df	MS
Model	278291.002	4	69572.7506
Residual	447167.048	1,818	245.966473
Total	725458.05	1,822	398.16578

Number of obs	=	1,823
F(4, 1818)	=	282.85
Prob > F	=	0.0000
R-squared	=	0.3836
Adj R-squared	=	0.3823
Root MSE	=	15.683

math4	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lexppp	21.84406	4.324429	5.05	0.000	13.36269	30.32543
lunch	-.4633164	.0143219	-32.35	0.000	-.4914054	-.4352274
enroll	.0196432	.0093826	2.09	0.036	.0012413	.0380451
expend	-6.25e-06	1.80e-06	-3.47	0.001	-9.79e-06	-2.71e-06
_cons	-91.46384	37.0014	-2.47	0.014	-164.0336	-18.89412

The estimated equation for Model 4 is as follows:

$$math4 = -91.464 + 21.844 * log(exppp) - .463 lunch + .020 enroll - (6.25 * 10^{-6}) expend$$

In Model 4, the coefficient for $log(exppp)$ is much larger than it is in our previous models. The model indicates that a 1% increase in expenditures per pupil increases the percentage of children satisfactory in mathematics by 21.844%, which is several times larger than it is in Models 2 and 3. This occurred because we removed *read4* from our analysis and because the intercept is now much greater in (negative) magnitude than it was in Model 2.

Furthermore, we see that all other variables have increased in magnitude by the removal of *read4*. The coefficient for *lunch* has quadrupled, the coefficient for *expend* has almost tripled, and the coefficient for *enroll* is twenty times greater. Still, all the signs for these variables are the same. Overall, *read4*

However, the R-squared value and adjusted R-square value are both 0.38. This means that this model is able to explain much less of the variation than our previous models. However, we removed a highly predictive independent variable in *read4*, so this is expected.

<u>*Model 5:*</u>

Model 5 is analogous to Model 3, except that *read4* is not included. Model 5 is represented as:

$$math4 = B_0 + B_1 * log(exppp) \; + B_2 lunch \; + B_3 lexpend + u$$

Our STATA output is listed in Figure 6:

Figure 6 – Stata Output for Model 5

```
. regress math4 lexppp lunch lexpend
```

Source	SS	df	MS			
				Number of obs	=	1,823
				F(3, 1819)	=	371.29
Model	275520.773	3	91840.2577	Prob > F	=	0.0000
Residual	449937.277	1,819	247.354193	R-squared	=	0.3798
				Adj R-squared	=	0.3788
Total	725458.05	1,822	398.16578	Root MSE	=	15.727

math4	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lexppp	13.37372	1.799705	7.43	0.000	9.844018	16.90343
lunch	-.4713807	.0143046	-32.95	0.000	-.4994359	-.4433254
lexpend	-4.839606	.9176376	-5.27	0.000	-6.63934	-3.039872
_cons	46.19535	17.91481	2.58	0.010	11.05959	81.3311

The estimated equation for Model 5 is as follows:

$$math4 = 46.195 + 13.374 * log(exppp) - .471 lunch - 4.840 lexpend$$

The coefficient for *lexppp* here is more moderate than it is in Model 4 but still much greater than it is in Models 2 and 3. Clearly, removing the *read4* variable made the impact of *lexppp* much more significant in general. We see, however, that the R-squared is the same for both Models 4 and 5. It appears that removing *enroll* from our equation did not reduce the percentage of the variation we can account for with these independent variables. Moreover, we see that the magnitudes of the other variables have increased as well in Model 5 when compared to Model 3. This reflects the same trend that we saw between Models 2 and 4. Overall, removing the *read4* variable from our dataset has made our models much less predictive of the variation but has increased the relative effect each independent variable has on PSM.

Model 6:

Our final model eliminates all of our secondary independent variables except for *lunch*. We are interested to see whether our results will have similar R-squared values and coefficients when compared to the two previous models. Model 6 is represented as:

$$math4 = B_0 + B_1 * log(exppp) \ + B_2 lunch + u$$

Our STATA output is listed in Figure 7:

Figure 7 – Stata Output for Model 6

```
. regress math4 lexppp lunch

      Source |       SS           df       MS      Number of obs   =     1,823
-------------+----------------------------------   F(2, 1820)      =    535.14
       Model |  268640.64         2  134320.32     Prob > F        =    0.0000
    Residual |  456817.41     1,820  250.998577    R-squared       =    0.3703
-------------+----------------------------------   Adj R-squared   =    0.3696
       Total |  725458.05     1,822  398.16578     Root MSE        =    15.843

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lexppp |   11.33798   1.770726     6.40   0.000     7.865115    14.81085
       lunch |  -.4712446   .0144096   -32.70   0.000    -.4995057   -.4429834
       _cons |  -6.342395   14.99906    -0.42   0.672    -35.75957    23.07478
------------------------------------------------------------------------------
```

The estimated equation for Model 6 is as follows:

$$math4 = -6.342 + 11.338 * log(exppp) - .471 lunch$$

In this model, a 1% change in EPS predicts an 11.337% increase in the percentage of students satisfactory in fourth grade mathematics. This is very slightly lower than the value we saw in Model 5 but still quite similar in magnitude. We also observe that the coefficient for the *lunch* variable did not change at all from Model 5. The intercept is the only element that significantly changed between the models. Still, our R-squared value in this model is very similar to Models 5 and 6; dropping *enroll, expend,* and *lenroll* did little to change the explained variation in the model.

Table 4 – Regression Results Summary

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*	*Model 6*
lexppp	-1.52 (2.17)	6.81** (2.92)	3.39*** (1.22)	21.84*** (4.32)	13.37*** (1.80)	11.34*** (1.77)
read4		0.78*** (0.02)	0.79*** (0.02)			

lunch		-0.11*** (0.01)	-0.11*** (0.01)	-0.46*** (0.01)	-0.47*** (0.01)	-0.47*** (0.014)
enroll		0.01 (0.01)		0.20** (0.01)		
expend		0.00* 0.00		0.00*** (0.00)		
lexpend			-1.27** (0.62)		-4.84*** (0.92)	
Intercept	84.84 (18.56)	-27.49 (24.85)	18.50 (12.00)	-91.46 (37.00)	46.20 (17.91)	-6.34 (15.00)
R-Squared	0.00	0.72	0.72	0.38	0.38	0.37

*Significant at 10%, **5%, ***1%

Table 4 provides a summary of our regression results. The significance levels confirm some of our analysis from above. In Model 1, we can see that *lexppp* is not significant at even a 10% level. *lexppp, read4,* and *lunch* are significant at a 1% level in Models 2-6 when they are included. They appear to be the variables most important to our linear regression models. All other independent variables were not always significant at the 1% level in our models.

V. Extensions

After making Models 4 and 5, we conducted an F-test to determine what independent variables should make up our final model. However, the results were not very indicative because our sample size is very large. Still, we include our results below. Our hypotheses for Model 5 were as follows:

$$H_0: B_3 = 0$$

$$H_1: H_0 is\ false$$

Using the R-square values from Models 5 and 6, we find that the F-value is 27.847. The critical F-value at a 1% level for this test is 6.63, but our F-value greatly surpasses it. As a result, we fail to reject our null hypothesis. Similarly, we used $H_0: B_3 = 0, B_4 = 0$ as our null hypothesis for the F-test on Model 4, in which we tested whether *enroll* and *expend* are jointly significant. The F-value for this test is 19.613, and the critical F-value is 4.61. Once again, the F-value is much greater than the critical value, and we fail to reject the null hypothesis. Similar F-tests were conducted to test the significance of the *lunch* independent variable, but the F-values obtained were greater than 500. Since the F-value for *lunch* was greater than those of the other variables, we decided that Model 6 should include *lunch* rather than the other variables.

Given the results of the scatterplot in Figure 1, linear regression may not be the best-suited regression type for our data. As a result, we conducted a logistic regression based on Model 6 using Stata software. Although a perfect comparison does not exist, we will assume that we can compare the R-squared value from our linear regression model to the pseudo R-squared value of the logistic model. The R-squared value is substantially lower for our logistic model when compared to its counterpart in Model 6. In addition, the p-value is large for both *lexppp* and *lunch*; neither variable is statistically significant in this regression. Given these results, we predict that a linear model is more suited for our dataset than is a logistic model, but more evidence is needed for a legitimate conclusion.

Figure 8 – Stata Output for Logistic Model

```
. logit math4 lexppp lunch

Iteration 0:   log likelihood = -8.5079645
Iteration 1:   log likelihood = -7.2712465
Iteration 2:   log likelihood = -6.6714274
Iteration 3:   log likelihood = -6.6520224
Iteration 4:   log likelihood = -6.6519795
Iteration 5:   log likelihood = -6.6519795

Logistic regression                             Number of obs     =      1,823
                                                LR chi2(2)        =       3.71
                                                Prob > chi2       =     0.1563
Log likelihood = -6.6519795                     Pseudo R2         =     0.2181

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lexppp |   3.316435   2.769585     1.20   0.231    -2.111852    8.744722
       lunch |  -.0851277   .0636007    -1.34   0.181    -.2097828    .0395274
       _cons |  -15.12877   22.70815    -0.67   0.505    -59.63592    29.37838
------------------------------------------------------------------------------
```

VI. Conclusions

Our multiple regression models have supported our initial hypothesis that there is a significant, positive correlation between our dependent variable, PSM, and our primary independent variable, EPS. The simple linear regression model, however, indicated that an increase in PSM would result in a decrease in EPS, but it could explain very little of the variation in PSM across schools. Our multiple regression models showed that this was not true when all other independent variables were held constant. *lexppp* was shown to be significant at a 1% level with a large effect on *math4* in the regression models, and these models were able to explain a reasonable amount of variation in *math4.*

Furthermore, we found *read4* and *lunch* to be the next most important independent variables. Removing *read4* from our models resulted in our R-squared values from about 0.7 to 0.3. This significant decrease was expected because *read4* is very similar to *math4* because it is another measure of academic success. We believe that removing *read4* from the results would allow us to make more practical recommendations. *lunch* was significant at a 1% variable in each of our linear models, and we believe that this variable was key to our dataset. Past studies have included the *lunch* data, and we found that a model with just *lexppp* and *lunch* is able to account for a significant amount of the variation in *math4*.

Resultingly, we conclude that increasing expenditures per student can increase the percentage of students satisfactory in 4th grade math and the overall quality of education at a school. However, we must note that not all interpretation of the results must be met with some limitations of our study. Namely, not all of the classic linear model assumptions were met, and our R-squared values for models without the *read4* variable are not very high.

Our research can represent a base for several paths of future research. One path could include a similar study in which different independent variables are included and are able to increase the R-squared values in the models. However, to research the effect of EPS on quality of education more effectively, we suggest a study that tracks certain schools over time as their EPS increases. This would allow for more factors to be controlled, and the study would have more accurate conclusions. Overall, since education has such a large effect on the lives of all citizens, we must continue to investigate avenues in which education can be improved.

References

Briggs, Jason C. (2013). A quantitative study of the percentage of variance in student Terra Nova Assessment scores accounted for by the variance in school size and per pupil expenditure in Catholic elementary schools. ProQuest Dissertations Publishing.

Carpenter, Carl A., II. (2006). The relationship of public elementary school size, student achievement, and per pupil expenditures in South Carolina. ProQuest Dissertations Publishing.

Goal 4 | Education. (n.d.). Retrieved from https://sdgs.un.org/goals/goal4

Hanushek, Eric A, Ruhose Jens, and Woessmann, Ludger. (2015). Knowledge Capital and Aggregate Income Differences: Development Accounting for U.S. States. https://doi.org/10.3386/w21295

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., and Barmer, A. (2019). The Condition of Education 2019. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mutuku, Stephen, & Korir, Julius. (2019). Government Expenditure and Quality of Education: A Case of Public Primary Schools in Kenya. Modern Economy, 10(12), 2405–2429. https://doi.org/10.4236/me.2019.1012152

Quality Education: Why It Matters. (n.d.). Retrieved from http://cdn.worldslargestlesson.globalgoals.org/2017/03/4.pdf